Related to other papers in this special issue	4 (p40); 11 (p108); 8 (p78)
Addressing FAIR principles	F

Unique, Persistent, Resolvable: Identifiers as the Foundation of FAIR

Nick Juty[1†], Sarala M. Wimalaratne[2], Stian Soiland-Reyes[1], John Kunze[3], Carole A. Goble[1] & Tim Clark[4]

[1]Department of Computer Science, The University of Manchester, Oxford Road, Manchester M13 9PL, UK

[2]European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, CB10 1SD, UK

[3]California Digital Library, Oakland, California 94612-2901, USA

[4]Data Science Institute, University of Virginia, Charlottesville, VA 22903-1738, USA

Keywords: Identifiers; metadata; findability; FAIR data; data infrastructures

Citation: N. Juty, S.M. Wimalaratne, S. Soiland-Reyes, J. Kunze, C.A. Goble & T. Clark. Unique, persistent, resolvable: Identifiers as the foundation of FAIR. Data Intelligence 2(2020), 30–39. doi: 10.1162/dint_a_00025

ABSTRACT

The FAIR principles describe characteristics intended to support access to and reuse of digital artifacts in the scientific research ecosystem. Persistent, globally unique identifiers, resolvable on the Web, and associated with a set of additional descriptive metadata, are foundational to FAIR data. Here we describe some basic principles and exemplars for their design, use and orchestration with other system elements to achieve FAIRness for digital research objects.

† Corresponding author: Nick Juty (E-mail: nsjuty@gmail.com, ORCID: 0000-0002-2036-8350).

1. INTRODUCTION

The FAIR principles [1] are a model to guide the implementation of interoperable digital research artifacts which are fully discoverable and ultimately reusable in the scientific research ecosystem. These digital artifacts must be preserved long term as the basis of modern scientific research. However, recent publications suggest that up to 80% of scientific research data are lost within 20 years [2]. This estimate does not take into account the effects of practical unavailability of such data, when they have not been robustly and accessibly archived, even when they are not actually "lost". Data attrition blocks the ability to validate results and reuse information in science.

"FAIR" as an approach to reduce data attrition and support reuse has seen growing support in EU- and US-funded projects[①] and influential science policy recommendations [3].

This article focuses specifically on identifiers as required for practical implementation of the foundational "F" element of FAIR, the "Findability" of digital research artifacts. All the other FAIR principles ("A", "I", and "R") are based upon it. We cannot access, interoperate or reuse data without knowing how to identify them. And ultimately, all criteria of FAIR influence how we must provision identifiers.

FAIR data, software, and other digital research objects collectively support the validity of scientific research as communicated in the literature. Likewise, the scientific literature in turn supports FAIR data, when properly referenced, by describing the methods by which the data have been obtained.

2. IDENTIFIERS RESOLVABLE ON THE INTERNET

To be "Findable", digital objects must be uniquely specifiable, i.e., they must have unique names, and must be locatable, by association of the unique name with a protocol for retrieval. They must also have at least some associated descriptive metadata to assist discovery and verification of that object when the identifier is not known *a priori*. Recent work on scientific data "commons" architectures [4] and multi-commons "data ecosystem frameworks" [5], highlighted here, reinforces the FAIR Accessibility principle that identifiers must be paired with such descriptive metadata [6,7,8].

An *identifier* is a unique name given to an object, property, set or class. Digital identifiers, by which we mean ordered strings of characters, uniquely name digital resources. An identifier may sometimes be "read", as if its string revealed meaningful semantics, but fundamentally, the underlying semantics of an identifier is that it *means* the resource to which it refers. Table 1 shows four common FAIR identifier types, which we discuss in detail later in this article.

① See https://gdc.cancer.gov, https://www.fairsfair.edu, and https://fairplus-project.edu.

Table 1. Examples of FAIR digital identifiers.

Identifier Type	Example	Object Identified
DOI	https://doi.org/10.1038/sdata.2016.18	A digital object – the publication *Wilkinson et al. 2016, Sci Data 3,160018*
ARK	http://n2t.net/ark:/65665/3c2e39526-e0c3-41ae-be4f-07558a9458eb	A physical specimen in the Smithsonian collection – holotype of the insect *Z. wardiana*
Identifiers.org	https://identifiers.org/uniprot:P0DP23	A digital object – information on the protein Calmodulin-1
PURL	http://purl.org/dc/terms/publisher	An abstract term in the DCTERMS vocabulary – "publisher"

Identifiers have a scope within which they are verifiably unique. For digital artifacts this may be local scope within some particular data collection, as is the case with typical "accession numbers", the concept for which originated in museum, library and specimen collection practice and was extended to the biomedical sequence and literature databases.

In current best citation practice, identifiers must be globally unique and resolvable on the Internet, where digital identifiers are ordinary Web addresses, technically URIs, or Uniform Resource Identifiers (URIs)[9]. A URI is a compact sequence of characters uniquely identifying a resource. Resources may be physical things (e.g., a dog, building, microscope, star, and person), which are described by associated information), digital (e.g., a document, data set, catalog record, program executable, service endpoint, or the actual objects of interest), or abstract (e.g., a vocabulary term).

URIs begin with a scheme name which specifies the method used to resolve them, e.g., *http*, *urn*, *ftp*, LDAP, etc. In this article we will be speaking primarily of secure *http* URIs, which begin with "https://" [10], but using the term *http* URI to mean both secure and insecure variants. And we will discuss their resolution in the context of REST interfaces [11,12], the canonical way to access required metadata, including the resolution endpoints, for any persistent identifier.

An *http* URI gives a protocol for finding (resolving) and accessing the information resource to which it refers, that is, it is used to locate the resource. As noted, we find the "Accessible" requirement of FAIR is also embedded in secure *https* URI based identifiers.

3. IDENTIFIER PERSISTENCE AND INDIRECTION

The hosting of identified resources can change; publishing companies or data archives may merge, or be acquired, or the internal infrastructure at an institution may be revamped, modifying the presentation of a resource's URIs. This means there is a problem if identifiers are dependent upon local hostnames; this can be mitigated by *indirection* through central public resolvers. These provide resolution for the identifier URI by redirecting *http* GET requests [12] to a different *http* URI hosted by the information provider.

Identifiers may be made *indirect* by (Pattern 1) registering each one with a redirecting service (along with some metadata) as it is issued; or (Pattern 2) by registering a unique namespace, identifier pattern and resolver endpoint for the resource provider just once. In either case indirect identifiers become "persistable" in that changes to the endpoint URIs can be made invisible and non-disruptive to users either individually (Pattern 1 above) or in bulk (Pattern 2) when the provider hosting structure changes. **Pattern 1** allows us to retrieve core metadata elements for individual identifiers. **Pattern 2** allows us to retrofit global resolution and persistence behavior to systems that formerly assigned only locally unique identifiers such as database accession numbers.

4. EXEMPLAR PERSISTENT, RESOLVABLE IDENTIFIERS

A variety of services have been developed and made available in support of persistent identifiers, facilitating their consistent use/reuse, as well as global collaborative efforts to converge upon best practices in these areas. We highlight some exemplar services here, all of which comply with the recommendations in [13].

4.1 Digital Object Identifiers

A very well-known example of the **Pattern 1** approach (above) is the Digital Object Identifier (DOI) system [14]. DOIs are widely used in the publishing industry to identify documents, and as data set identifiers and for data citation and interoperability, besides other uses. They provide a canonical example of how Pattern 1 should be implemented.

The International DOI Foundation[②] maintains the DOI system, supported by its various domain-centric Registration Agencies (RAs). The most important RA for our purposes is DataCite[③], a member supported organization that registers data sets [15]. DataCite membership gives an organization the right to mint a certain number of DOIs per year, at various subscription levels. Over 16 million unique DOIs have been registered with DataCite. DOI resolution is free.

DOIs consist of a DOI name which is resolved at https://doi.org, with the full URI formulated according to the pattern https://doi.org/<DOI_name>. DOI names in turn consist of a *prefix* and a *suffix*, separated by a forward slash. The *prefix* is a code indicating the registrant who issues the DOI, e.g., Harvard University Dataverse – 10.7910; Dryad Digital Repository – 10.5061. The *suffix* is the identifier, in any form, assigned by the registrant.

② https://doi.org.
③ https://datacite.org.

Thus an *http* GET request to:

```
https://doi.org/10.5061/dryad.k5r2bb0
```

Redirects to the Dryad entry at this URI:

```
https://datadryad.org/resource/doi:10.5061/dryad.k5r2bb0
```

which returns metadata about the entry:

```
Data from: The Muensterelloidea: phylogeny and character evolution of Mesozoic stem octopods
Authors: Fuchs D, Iba Y, Heyng A, Iijima M, Klug C, Larson NL, Schweigert G
Date Published: June 28, 2019
```

including the resolution endpoint for its data package, here:

```
Data package: https://datadryad.org/bitstream/handle/10255/dryad.192323/Appendix%202_Data%20
matrix.nex?sequence=1
```

Metadata, including the data package URI, will have been registered with DataCite at the time the identifier was assigned. Full resolution to the actual data is a multi-step process. First, the client sends an *http* GET request to the doi.org resolver for the DOI of interest. Then doi.org sends a redirect message to the client with the address of the data set's "landing page", which holds the metadata, including the resolution endpoint(s) for its data.

Why is this multi-redirect process required? The first redirect separates the endpoint of the data from their persistent identifier, providing isolation from infrastructure changes at the provider side. The second redirect allows decisions to be made about how to handle the data set prior to access. For example, we may want to ensure that we have the correct data set by scanning the metadata; we may not want to directly download a large data set before we ascertain it is the correct one (via the metadata); and if very large, we may wish to consider transit costs, and transport speed, before we decide how to access it. If the data exist at multiple mirror sites, we might want to consider proximity. These decisions can be made in real time by humans working at a browser, or on a command line; or they can be made algorithmically by "content negotiation" between services [16,17].

Metadata at the DataCite service is provided in both human and machine readable forms. Machine readable metadata is provided both in DataCite XML and in the schema.org [18] vocabulary, serialized as JSON-LD [19]. Future work may also enable incorporation of terms from domain-specific schema.org extensions such as bioschemas[⑧] [20].

⑧ https://bioschemas.org.

4.2 Archival Resource Keys

Archival Resource Keys (ARKs) are another widely used persistent identifier, supported by the California Digital Library [21], in collaboration with DuraSpace[5]. ARKs work similarly to DOIs, but are more permissive in design. Over 500 registered organizations have created over 3.2 billion ARKs. There is no fee for registering or resolving ARKs. Metadata is strongly recommended for publicly released ARKs, but is optional and flexible while the resource is being planned, reviewed, tested, etc.

The ARK scheme is decentralized; any organization can run its own resolver, but all can be resolved through the global Name-to-Thing resolver[6]. N2T operates in two modes. It stores records for millions of ARKs registered by the EZID service[7] [22] and the Internet Archive. And in collaboration with identifiers.org (below), it stores 3,500 rules for redirecting identifiers that it does not store individually.

ARKs have the syntax:
https://NMA/ark:/NAAN/Name[Qualifier]
where:

- NMA is the Name Mapping Authority;
- NAAN is the Name Assigning Authority Number (like a DOI prefix),
- Name is the local identifier issued by the NAAN, and
- [Qualifier] is an optional qualifier.

Name Mapping Authorities operate ARK resolvers. Name Assigning Authorities have NAANs allowing them to mint globally unique ARKs.

4.3 Identifiers.org and CURIEs

The identifiers.org system [23,24,25] has been providing globally resolvable *http* URI-based identifiers to the scientific community for over a decade. The core system consists of a resolver and a namespace mapper, following Pattern 2 (Section 3). It registers a unique namespace, a namespace prefix, an identifier regex and a local identifier resolver URI, to provide globally unique persistent identifiers. It records metadata only for the identifier namespace, not for individual identifiers, relying upon the data provider to expose the appropriate metadata within the resolved landing page.

Identifiers.org URIs follow the syntax:
https://Resolver/[Provider/]Namespace:Identifier
where

[5] https://ARKsInTheOpen.org.
[6] https://N2T.net/.
[7] https://ezid.cdlib.org.

- Resolver is resolution host, currently either identifiers.org or n2t.net;
- Provider is an *optional* term indicating specific providers for a Namespace, e.g., "RCSB", "PDBe", "PDBj", etc., for the Protein Data Bank's various providers;
- Namespace is a registered prefix assigned to an identifier issuing authority; and
- Identifier is the locally assigned identifier.

Taken together, a namespace prefix and colon followed by a local identifier constitute a CURIE, or Compact Identifier [13,23].

Identifiers.org records all known resource-specific access URIs for a data set or data record, including alternative access URIs. This information allows numerous equivalent URIs to be considered to be a single canonical URI string. This *mapping*, retrievable as a data set in its own right, is integral to platforms like EMBL-EBI RDF and Open PHACTS [26], where identifiers.org URIs act as a semantic glue in distributed queries.

4.4 PURLs

Persistent Uniform Resource Locators (PURLs) generically are *http* URIs whose services redirect GET requests to another location-based URI, with the added promise of long-term *persistence*. All identifier types discussed above meet this definition. However, in common usage, a PURL is an identifier minted and resolved at https://purl.org, hosted since 2016 by the Internet Archive[8], after being transferred from OCLC, where the service started in 1995. That the ownership was transferred to the Internet Archive highlights that particular care was taken to preserve and continue operation of this identifier infrastructure and its registrations. A separate PURL registry exists at https://purl.obolibrary.org, hosted by the OBO Foundry. OBO Library PURLs are widely used by the OBO ontologies community to persistently reference their information artifacts, for which they are well suited[9].

Users of purl.org register to gain control of a sub-path, e.g., http://purl.org/dc/ and then assign and modify the redirection targets for individual entries like http://purl.org/dc/terms/. The ability to update the PURL target is crucial for persistence, for instance if ownership of a service changes, or the server infrastructure changes URL layout. In this sense a PURL is a counter-measure to unintended violations of "Cool URIs don't change" [27].

5. DATA CITATION

The Joint Declaration of Data Citation Principles (JDDCP) provides a set of guiding principles for publishing and citing research data. Many implementation guidelines developed for JDDCP [15, 16, 23, 28, 29] provide rich foundational practices for FAIR, including globally unique persistent identifiers, indirect

[8] https://archive.org.

[9] However, we do not recommend PURLs in general as primary identifiers for persistently archived FAIR data.

resolution to a landing page or service, and accessibility of human and machine readable metadata. In our view, it is not possible to reliably reuse data without robust links to provenance documentation. Ultimately, literature describing how data were obtained provides the rationale for accepting them as reliable. Conversely, no assertion in the scientific literature is ultimately credible unless backed by reliable data. Good data sharing, archiving and citation practices followed by authors, publishers, repositories, funders and institutions, along with good identifier and metadata practices, and appropriate credit incentives [30,31], will promote creation of normative FAIR data integrally associated with the literature.

6. IDENTIFIERS PLUS METADATA

Sensible strategies that work across projects, domains and infrastructures permit realization of an open landscape where science can operate reliably with massive amounts of data. Large scale "data commons" projects such as the European Open Science Cloud, the NCI Genomic Data Commons, and the NHLBI Data Stage are initial steps in this direction. Investigators in such programs understand that frameworks for interoperating across commons in a "data ecosystem" [5] of FAIR research objects require not only persistent resolvable identifiers, but well-documented APIs for associated human- and machine-readable metadata. These metadata should optimally include not only provenance information but descriptive "guide" metadata to enable discoverability. A large, active community is engaged in determining and establishing best practices around persistent identifiers, metadata standards and exposure mechanisms; there are many active "interest" and "working" groups engaged in such activities under the auspices of the Research Data Alliance®.

Properly managed persistent identifiers with their associated metadata are the primary access point to enable **Findability** in the FAIR data ecosystem, and to provide a basis for all the other FAIR attributes.

AUTHOR CONTRIBUTIONS

N. Juty (nick.juty@manchester.ac.uk) conceived, organized and wrote initial drafts of this article, and managed the discussion of drafts. S. Wimalaratne (sarala.dissanayake@gmail.com), S. Soiland-Reyes (soiland-reyes@manchester.ac.uk), J. Kunze (jak@ucop.edu), C. Goble (carole.goble@manchester.ac.uk) and T. Clark (twc8q@virginia.edu) contributed discussion material in preparing the draft. T. Clark, J. Kunze, and N. Juty co-wrote the final draft. All authors reviewed and edited the final manuscript.

ACKNOWLEDGEMENTS

This work was supported in part by the European Union's Horizon 2020 program under grant agreements 777523, FREYA, "Connected Open Identifiers for Discovery, Access and Use of Research Resources", 654248, CORBEL, "Coordinated Research Infrastructures Building Enduring Life-science services", and

® https://www.rd-alliance.org/.

823830, Bioexcel2, "BioExcel-2 Centre of Excellence for Computational Biomolecular Research". Many thanks to Paul Groth for his helpful comments on the manuscript.

REFERENCES

[1] M.D. Wilkinson, M. Dumontier, Ij.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, … & B. Mons. The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3(2016), Article No. 160018. doi: 10.1038/sdata.2016.18.

[2] T.H. Vines, A.Y.K. Albert, R.L. Andrew, F. Débarre, D.G. Bock, M.T. Franklin, … & D.J. Rennison. The availability of research data declines rapidly with article age. Current Biology 24(1) (2014), 94–97. doi: 10.1016/j.cub.2013.11.014.

[3] Committee on Toward an Open Science Enterprise. Open science by design: Realizing a vision for 21st century research. Washington, DC: National Academies Press, 2018. isbn: 978-0-309-47624-9.

[4] T. Clark & D.S. Katz (eds.). DCPPC DRAFT: KC2 globally unique identifier services. NIH Data Commons Pilot Project Consortium (2018). Available at: https://public.nihdatacommons.us/DCPPC-DRAFT-8_KC2/.

[5] R.L. Grossman. Data lakes, clouds, and commons: A review of platforms for analyzing and sharing genomic data. Trends in Genetics 35(3)(2019), 223–234. doi: 10.1016/j.tig.2018.12.006.

[6] T. Weigel, U. Schwardmann, J. Klump, S. Bendoukha & R. Quick. Making data and workflows findable for machines. Data Intelligence 2(2020), 40–46. doi: 10.1162/dint_a_00026.

[7] C. Goble, S. Cohen-Boulakia, S. Soiland-Reyes, D. Garijo, Y. Gil, M.R. Crusoe, K. Peters & D. Schober. FAIR computational workflows. Data Intelligence 2(2020), 108–121. doi: 10.1162/dint_a_00033.

[8] P. Groth, H. Cousijn, T. Clark & C. Goble. FAIR data reuse – the path through data citation. Data Intelligence 2(2020), 78–86. doi: 10.1162/dint_a_00030.

[9] T. Berners-Lee, R. Fielding & L. Masinter. IETF RFC 3986: Uniform Resource Identifier (URI): Generic Syntax, 2005. Available at: http://tools.ietf.org/html/rfc3986.

[10] R. Fielding, J. Getty, J. Mogul, H. Frystyk, L. Masinter, P. Leach & T. Berners-Lee. Hypertext transfer protocol — HTTP/1.1, IETF RFC 2616, Internet Engineering Task Force. (1999). Available at: http://www.ietf.org/rfc/rfc2616.txt.

[11] R.T. Fielding. Architectural styles and the design of network-based software architectures. PhD dissertation, University of California, 2000. Available at: http://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf.

[12] R.T. Fielding & R.N. Taylor. Principled design of the modern web architecture. ACM Transactions on Internet Technology 2(2) (2002), 115–150. doi: 10.1145/514183.514185.

[13] J.A. McMurry, N. Juty, N. Blomberg, T. Burdett, T. Conlin, N. Conte, … & H. Parkinson. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLoS Biology 15(6)(2017), e2001414. doi: 10.1371/journal.pbio.2001414.

[14] ISO/TC46, Information and documentation—Digital object identifier system. International Standards Organization, 2012. Available at: https://www.iso.org/obp/ui/#iso:std:iso:26324:en.

[15] J. Brase. DataCite—A global registration agency for research data. In: Proceedings of the 2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, 2009, pp. 257–261. doi: 10.1109/COINFO.2009.66.

[16] M. Fenner, M. Crosas, J.S. Grethe, D. Kennedy, H. Hermjakob, P. Rocca-Serra, … & T. Clark. A data citation roadmap for scholarly data repositories. Scientific Data 6 (2019), Article No. 28. doi: 10.1038/s41597-019-0031-8.

[17] J. Starr, E. Castro, M. Crosas, M. Dumontier, R.R. Downs, R. Duerr, … & T. Clark. Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science 1 (2015), e1. doi: 10.7717/peerj-cs.1.
[18] R.V. Guha, D. Brickley & S. Macbeth. Schema.org: Evolution of structured data on the web. Communications of the ACM 59(2)(2016), 44–51. doi: 10.1145/2844544.
[19] M. Sporny, G. Kellogg, M. Lanthaler, D. Longley & N. Lindström. JSON-LD 1.0: A JSON-based serialization for linked data, W3C Recommendation 16 January 2014. (2014). Available at: http://www.w3.org/TR/2014/REC-json-ld-20140116/.
[20] Bioschemas Community. Bioschemas specifications (2018). Available at: https://bioschemas.org/specifications.
[21] J. Kunze & R. Rodgers. The ARK identifier scheme (2008). Available at: https://escholarship.org/uc/item/9p9863nc.
[22] University of California. The EZID API, Version 2, (2019). Available at: https://ezid.cdlib.org/doc/apidoc.html.
[23] S.M. Wimalaratne, N. Juty, J. Kunze, G. Janée, J.A. McMurry, N. Beard, … & T. Clark. Uniform resolution of compact identifiers for biomedical data. Scientific Data 5(2018), Article No. 180029. doi:10.1038/sdata.2018.29.
[24] N. Juty, N. Le Novère & C. Laibe. Identifiers.org and MIRIAM Registry: Community resources to provide persistent identification. Nucleic Acids Research 40(D1)(2012), D580–D586. doi: 10.1093/nar/gkr1097.
[25] C. Laibe & N. Le Novère. MIRIAM Resources: Tools to generate and resolve robust cross-references in Systems Biology. BMC Systems Biology 1 (2007), 58. doi: 10.1186/1752-0509-1-58.
[26] A.J. Williams, L. Harland, P. Groth, S. Pettifer, C. Chichester, E.L. Willighagen, … & B. Mons. Open PHACTS: Semantic interoperability for drug discovery. Drug Discovery Today 17(21-22)(2012), 1188–1198. doi:10.1016/j.drudis.2012.05.016.
[27] T.B. Lee. World Wide Web Consortium (W3C). Cool URIs don't change. Available at: https://www.w3.org/Provider/Style/URI.
[28] Data citation needed (Editorial). Scientific Data 6 (2019), Article No. 27. doi: 10.1038/s41597-019-0026-5.
[29] H. Cousijn, A. Kenall, E. Ganley, M. Harrison, D. Kernohan, T. Lemberger, … & T. Clark. A data citation roadmap for scientific publishers. Scientific Data 5 (2018), Article No. 180259. doi: 10.1038/sdata.2018.259.
[30] B.E. Bierer, M. Crosas & H.H. Pierce. Data authorship as an incentive to data sharing. The New England Journal of Medicine 376(17)(2017), 1684–1687. doi: 10.1056/NEJMsb1616595.
[31] H.H. Pierce, A. Dev, E. Statham & B.E. Bierer. Credit data generators for data reuse. Nature 570 (2019), 30–32. doi: 10.1038/d41586-019-01715-4.